QUANTUM LEAP
INNOVATIONS

March 4, 2011

COL James McMains
Program Manager, Code 30
Office of Naval Research
875 North Randolph Street
Arlington, VA 22203-1995
james.mcmains@navy.mil

RE: Contract N00014-05-C-0541 - Final Report

Dear COL McMains,

Work under contract N00014-05-C-0541 has been completed. Attached please find our Final Report and the SF-298 Report Documentation Page for:

Integrated Asymmetric Urban Warfare Program (IAUWP) - Code 30

Covering the period January 2006 - January 2011

I will provide 2 CD's containing the software (CLIN 0002 and 0004) developed for this contract via Federal Express.

Thank you for your assistance on the above noted program. Copies have been distributed as per the Contract Data Requirements List – Instructions for Distribution.

Sincerely,

Frank Abbott

Frank T. Abbott
VP of Finance, CFO
fta@quantumleap.us

cc: Dr. Ganesh Vaidyanathan, Project Manager, Code 30, QLI gv@quantumleap.us
Administrative Contracting Officer – Stanley Brown, stanley.brown@dcma.mil
Director, Naval Research Lab, Attn Code 5596, reports@library.nrl.navy.mil
Defense Technical Information Center, tr@dtic.mil

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)* 03-03-2011	**2. REPORT TYPE** Final Report	**3. DATES COVERED** *(From - To)* Jan 2006 - Jan 2011

4. TITLE AND SUBTITLE
Integrated Asymmetric Urban Warfare Program (IAUWP)

5a. CONTRACT NUMBER
N00014-05-C-0541-P00002

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
0602131M

6. AUTHOR(S)
Abbott, Franklin T.
Vaidyanathan, Ganesh

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Quantum Leap Innovations, Inc.
3 Innovation Way, Suite 100
Newark, DE 19711-5456

8. PERFORMING ORGANIZATION REPORT NUMBER
QLI-TR-11-001

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Office of Naval Research
ONR Code 30
875 North Randolph Street
Arlington, VA 22203-1995

10. SPONSOR/MONITOR'S ACRONYM(S)
ONR

11. SPONSORING/MONITORING AGENCY REPORT NUMBER

12. DISTRIBUTION AVAILABILITY STATEMENT
Approved for public release; distribution unlimited. 03 March 2011. Other requests for this document shall be referred to the Office of Naval Research, COL James McMains, Code 30, 875 N. Randolph St, Arlington, VA 22203-1995

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The aim of the IAUWP is to develop innovative technology that can be deployed in support of current and future military operations that include new risks from asymmetrical threats. The program includes the development, prototyping and testing of innovative intelligent computing software technologies which are targeted at supporting the warfighter in the urban and asymmetric battlespace. The QLI program emphasizes intelligent computing technologies that address current and future evolving threats to our forces. One early goal of the IAUWP is to explore the use of innovative technologies under the ONR-initiated concept of operations – Operational Adaptation (OA). OA creates the capability to develop and sustain a decision/action tempo that is beyond an irregular/terrorist adversary's ability to maintain. OA seeks to counter the effects of adversaries employing asymmetric tactics against US forces. In this report, we describe a novel Predictive Analytics technology (Flexscape) that enables data driven hypothesis generation and testing. An ensemble of Bayesian network models are automatically generated from data that can then be used to answer "what-if" questions to generate optimal responses within dynamic data environments. We further report on the analysis of data gathered during two ONR training exercises. QLI's Predictive Analytics software has been used to examine many diverse publicly available datasets during the development and proof of concept stages.

15. SUBJECT TERMS
Data Analytics, Hypothesis Generation, Predictive Analytics, Sensor Data Analysis, Automatic Model Building, Course of Action Planning, Command and Control, Knowledge Discovery, Collaboration

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT** UU	**18. NUMBER OF PAGES** 19	**19a. NAME OF RESPONSIBLE PERSON** Dr. Ganesh Vaidyanathan
a. REPORT Unclassified	**b. ABSTRACT** Unclassified	**c. THIS PAGE** Unclassified			**19b. TELEPONE NUMBER (*Include area code)*** 302-894-8044

Quantum Leap Innovations, Inc.
Delaware Technology Park
3 Innovation Way, Suite 100
Newark, DE 19711

QLI-TR-11-001
January 2011

Integrated Asymmetric Urban Warfare Program (IAUWP)

Final Report

ONR Contract N00014-05-C-0541

Sponsored by the Office of Naval Research, Arlington, VA 22203

Abstract

The aim of the IAUWP is to develop innovative technology that can be deployed in support of current and future military operations that include new risks from asymmetrical threats. The program includes the development, prototyping and testing of innovative intelligent computing software technologies which are targeted at supporting the warfighter in the urban and asymmetric battlespace. The QLI program emphasizes intelligent computing technologies that address current and future evolving threats to our forces. One early goal of the IAUWP is to explore the use of innovative technologies under the ONR-initiated concept of operations – Operational Adaptation (OA). OA creates the capability to develop and sustain a decision/action tempo that is beyond an irregular/terrorist adversary's ability to maintain. OA seeks to counter the effects of adversaries employing asymmetric tactics against US forces. In this report, we describe a novel Predictive Analytics technology (Flexscape) that enables data driven hypothesis generation and testing. An ensemble of Bayesian network models are automatically generated from data that can then be used to answer "what-if" questions to generate optimal responses within dynamic data environments. We further report on the analysis of data gathered during two ONR training exercises. QLI's Predictive Analytics software has been used to examine many diverse publicly available datasets during the development and proof of concept stages.

Contents

List of Figures

List of Tables

1. Summary

Executive Summary

This final technical report summarizes Quantum Leap Innovations' (QLI) accomplishments with the Integrated Asymmetric Urban Warfare Program (IAUWP) through the contract close date of January 11, 2011 on ONR Contract N00014-05-C-0541-P00002. Having secured clearance status for key technical personnel, QLI focused our final efforts on performing analysis on classified Office of Naval Research (ONR) Code 30 datasets and preparing final reports and paperwork.

Summary of Accomplishments

Throughout the period of performance on this contract, QLI has been working on development and testing of components of the LeapWorks® Data Analytics platform. In particular, the LeapWorks Predictive Analytics component (formerly named Flexscape™) provides the capability to identify complex relationships inherent within a dataset. Models are built directly from the vast amounts of available data generating more accurate, useable and flexible models ready for advanced data analytics.

In addition to available ONR Code 30 data, QLI continues to actively seek appropriate datasets elsewhere to further validate the Flexscape technology. Examples of alternative datasets are described as use cases. QLI continues to investigate a variety of datasets including those in healthcare, pharmaceutical, financial, and consumer trends.

2. Motivation from the SOW

The aim of the IAUWP is to develop innovative technology that can be deployed in support of current and future military operations that include new risks from asymmetrical threats. The program intends to generate technologies that enhance US military operations in the field and seek to mitigate the impact of enemy actions on mission critical capabilities. The program includes the development, prototyping and testing of innovative intelligent computing software technologies which are targeted at supporting the warfighter in the urban and asymmetric battlespace.

In order to ensure relevance and to shorten the time from development to transitioned capability the work is to be fully integrated with ONR's Expeditionary Maneuver Warfare & Combating Terrorism Department (Code 30).

The work performed under IAUWP is focused on applying the resources of QLI to high-payoff deliverables that support the goals of the Code 30 programs in:

> "*develop(ing) future combat capabilities for Naval Expeditionary Maneuver Warfare and the Department's role in Combating Terrorism through the exploitation and subsequent application of Science and Technology in order to enhance the ability of the Navy-Marine Corps team to achieve assured access and conduct decisive operations as the naval portion of a Joint campaign."*

3. Background

Quantum Leap Innovations is a technology company developing and deploying software products focused on the transformation of data into information and information into knowledge. Our software addresses problems that are characterized as being complex and dynamic. Our work is based on distributed computing, intelligent agents, and automated knowledge discovery technologies.

The QLI program emphasizes intelligent computing technologies that address current and future evolving threats to our forces. By linking the program to current efforts that are underway at ONR and related organizations, the IAUWP remains tightly coupled to enduser defined requirements. This also ensures that the technologies developed by QLI are readily able to transition into service use. This is important, given the evolving tactics of enemy forces that are operating in an asymmetric battlespace. One early goal of the IAUWP is to explore the use of innovative technologies under the ONR-initiated concept of operations – Operational Adaptation (OA). OA creates the capability to develop and sustain a decision/action tempo that is beyond an irregular/terrorist adversary's ability to maintain. It is key to seizing and holding the initiative, and maintaining a dominant position of power against irregular threats. OA seeks to counter the effects of adversaries employing asymmetric tactics against US forces.

1. In this report, we describe a novel Predictive Analytics technology (Flexscape) developed under IAUWP that enables data driven hypothesis generation and testing. An ensemble of Bayesian network models are automatically generated from data that can then be used to answer "what-if" questions to generate optimal responses within dynamic data environments. For example, the data might represent sensor data that might indicate an imminent threat. The advantage of using Bayesian network models is that incomplete information can be assessed by the models to both predict possible outcomes as well as to generate hypotheses around the most likely outcomes. These capabilities are synergistic with the core themes of OA:

   > *The ability to gain, maintain, or recover the tactical, operational, and strategic initiative over an irregular threat by anticipating threat measures/ countermeasures and by facilitating the dynamic tailoring of friendly forces, capabilities, actions, and TTPs to defeat these measures/countermeasures. OA provides a capability to dynamically understand what decisions US forces can make or cause the enemy to make in order to generate an exploitable spatial, temporal, physical, or psychological advantage.*

2. We further report on the analysis of data gathered during two ONR training exercises. QLI's Predictive Analytics software has been used to examine many diverse publicly available datasets during the development and proof of concept stages. At the core, this technology is designed to build models directly from the data. QLI's experience with ONR Code 30 data is outlined in this document.

4. Contract Activities

4.1 Flexscape – Data Driven Hypothesis generation and testing

Summary:

The present approach relates to a method for generating hypotheses automatically from graphical models built directly from data. The method of the present approach links three key scientific concepts to enable hypothesis generation from data driven hypothesis-models:

1. Use of information theory based measures to identify informative feature subsets within the data.
2. Automatic generation of graphical models from the informative data subsets identified from step 1.
3. Application of optimization methods to graphical models to enable hypothesis generation.

The integration of these three concepts can enable scalable approaches to hypothesis generation from large, complex data environments. The use of graphical models as the model representation can allow prior knowledge to be effectively integrated into the modeling environment.

Background:

Hypothesis generation and testing has long been a cornerstone for the scientific method. The traditional scientific process has been to perform experiments to gather data. The data is then analyzed and human expertise is used to explain the data in the form of scientific principles that act both as an effective data compression mechanism as well as a means for generating new hypotheses that can be tested. More recently, with the rapid growth in data collection and the development of new data analysis methods, the question of whether the traditional scientific process can be facilitated through automation has become increasingly important.

The method of the present approach uses data to automatically build “hypothesis-models” which can be used to test and generate hypotheses. A hypothesis may be viewed as a “control strategy” aimed at achieving a desired result. For example, in a health care/life sciences context, a hypothesis can represent a preferred combination of treatments to mitigate the future impact of a disease. In a manufacturing context, a hypothesis can represent a set of process conditions that can optimize desired product properties. In a financial context, a hypothesis can represent a trading strategy to maximize profits. In the method of the present approach, a hypothesis thus represents a set of actions that can be taken in order to achieve a desired result with high probability. An important element of the present approach is to generate one or more hypotheses directly from data through the analysis of automatically generated hypothesis-models.

The method of the present approach links three key scientific concepts to enable hypothesis generation from data driven hypothesis-models:

1. Use of information theory based measures to identify informative feature subsets within the data.
2. Automatic generation of graphical models from the informative data subsets identified from step 1.
3. Application of optimization methods to graphical models to enable hypothesis generation.

The integration of these three concepts can enable scalable approaches to hypothesis generation from large, complex data environments. The use of graphical models as the model representation can allow prior knowledge to be effectively integrated into the modeling environment.

Furthermore, the method of the present approach extends the concepts outlined above to time varying data environments to enable both a forecasting capability as well as dynamic risk management strategies. In this instance, the graphical models encode temporal associations across the data, and the application of optimization methods on these dynamical graphical models results in prognostic hypotheses with associated uncertainties. Dynamic control strategies in a probabilistic data environment can be used to both anticipate and respond proactively to imminent threats that are intrinsic to operational adaptation.

Prior Art:

Bayesian networks are probabilistic graphical models that represent a set of random variables and their conditional independencies via a directed acyclic graph (DAG). The transparency of Bayesian networks enables the representation of hierarchical relations between variables through parent-child linkages.

- Pearl, Judea. *Causality: Models, Reasoning, and Inference*. Cambridge University Press (2000). ISBN 0-521-77362-8.

There is extensive literature relating to the learning of Bayesian networks directly from data including:

- Heckerman, David (March 1, 1995). "Tutorial on Learning with Bayesian Networks". in Jordan, Michael Irwin. *Learning in Graphical Models*. Adaptive Computation and Machine Learning. Cambridge, Massachusetts: MIT Press. 1998. pp. 301–354. ISBN 0-262-60032-3.
- Neapolitan, R.E. *Learning Bayesian Networks*, Prentice Hall, Upper Saddle River, NJ, 2004.

Structure learning methods such as the well known K2 algorithm assume a hierarchical ordering of variables to guide the learning.

- The well known K2 algorithm, Cooper, G.F. and Herskovits, E. (1992)
- A Bayesian method for the induction of probabilistic networks from data. *Mach. Learn*, 9, 309–347.)

Faulkner has described heuristic methods for finding optimal variable ordering to guide structure learning ("K2GA: Heuristically Guided Evolution of Bayesian Network Structures from Data", Faulkner, E., Proceedings of the IEEE Symposium of Computational Intelligence and Multi Criteria Decision Making, Honolulu HI, April 1-5, 2007). However, as Bostwick et al. have discussed, "the entire prior hypothesis space for even a moderately large relational database is so large that any Bayesian network attempting to capture it would be computationally intractable. (For example, some nodes would have tens or hundreds of thousands of states).” (CADRE: A System for Abductive Reasoning over Very Large Datasets. Daniel F. Bostwick, D. B. Hunter, N. J. Pioch. 2006. www.aaai.org/).

Yuan et al. discuss a general framework for generating multivariate explanations in Bayesian networks. However, they do not discuss the automatic generation of Bayesian networks from data to drive their explanation framework (Yuan, C. and Lu, T.C. A General Framework for Generating Multivariate Explanations in Bayesian Networks. Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence (2008) pp 1119-1124). Hypothesis generation associated with Bayesian networks has been primarily used in systems biology. Botstein et al. discuss the use of a “A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*)” where the role of data is primarily to provide evidence to Bayesian network models that have been constructed by domain experts rather than from the data (Troyanskaya, O.G., Dolinski, K., Owne, A.B., Altman, R.B. and Botstein, D. A. Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). PNAS July 8, 2003 vol. 100 no. 14 8348-8353). In the systems biology community, hypothesis generation from Bayesian networks has primarily been associated with the validation of linkages within a Bayesian network structure that has been postulated by domain experts:

- Weinreb, G.E., Kapustina, M.T., Jacobson K., Elston, T.C. In Silico Generation of Alternative Hypotheses Using Causal Mapping (CMAP). PloS ONE 4 (4): e5378. doi:10.1371/journal.pone.0005378, 2009.
- Rodin, A., Mosley, T.H., Clark, A.G., Sing, C.F. and Boerwinkle, E., Mining Genetic Epidemiology Data with Bayesian Networks Application to APOE Gene Variation and Plasma Lipid Levels, J. Comput. Biol.: 12 (1): 1-11, 2005.
- Pratt, D. R. et al., Causal Analysis in complex biological systems, US Patent No. 20070225956, issued September 27, 2007.

In US Patent No. 7,512,497 (Periwal, V., Systems and methods for inferring biological networks, issued March 31, 2009), optimization methods are used to infer cellular networks from a database of links. However, this patent does not teach how to generate the links database using information measures applied to raw data. In US Patent 6,941,287 (Vaidyanathan, A.G. et al., Distributed hierarchical evolutionary modeling and visualization of empirical data, issued September 6, 2005), Nishi entropy methods are used to identify informative features from data. However, Vaidyanathan et al. do not teach the automatic generation of Bayesian networks from the data. In addition, Vaidyanathan et al. do not teach the use of optimization methods applied to Bayesian networks to generate hypotheses.

In the present approach, a hypothesis is defined by a set of variable states that optimize a statistical measure associated with a desired outcome. The measure is computed using one or more Bayesian networks that have either been constructed directly from an informative data subset or that have been guided by an informative data subset. Further, the methods of the present approach alleviate the scalability difficulties by using information theory based feature reduction techniques to identify an informative subset of features using a mutual information measure. The reduced data set can be used by a structure learning algorithm such as the K2GA algorithm for efficient structure learning. One or more network structures can be learned from the data. The methods of the present approach further apply optimization methods on the informative Bayesian network structures to generate optimal hypotheses. The three key elements of the present approach: Information theory guided feature reduction, automated structure learning and automated hypothesis generation using optimization technologies provide the basis for scalable data driven hypothesis generation and testing.

The method of the present approach can also be extended to dynamical systems to provide a basis for dynamic risk management. In a dynamic environment, individual features can be extended into a list of (feature, time offset) feature pairs, where the time offset is measured against a reference time. The methods of the present approach can be used to analyze the extended dimensionality space covered by time stamped feature pairs to:

a. Reduce the dimensionality of the feature pair space using information theory based measures.
b. Sort the feature pairs in descending order so that the earlier time offsets occur earlier than the later time offsets.
c. Automatically generate at least one dynamic Bayesian network from the sorted data. Sorting the data as described will preserve the proper temporal sequencing between nodes within the network.
d. Apply optimization methods to at least one dynamic Bayesian network to generate a hypothesis.
e. Apply inference techniques on at least one dynamic Bayesian network to test a hypothesis.

The capability to generate a hypothesis from a data driven, dynamic Bayesian network can alleviate problems associated with classical time series analysis techniques such as ARIMA, recurrent neural networks and Monte Carlo Markov Chains which are difficult to employ in high dimensional data environments (Murphy, K. P., Dynamic Bayesian Networks: Representation, Inference and Learning, Ph.D. dissertation, University of California Berkeley, 2002).

The information theory based measures to reduce the dimensionality of the feature pair space can be used to zoom in on the most informative time lags to drive forecasts. In addition, the probabilistic nature of Bayesian networks can be used to calculate the uncertainty of the forecast that can be used as a basis for dynamic risk management in several domains, including financial services, health care and life sciences and manufacturing.

Summary of Approach:
Flexscape: Data Driven Hypothesis Testing and Generation System

The method of the present approach (Flexscape) uses data to automatically build “hypothesis-models” which can be used to test and generate hypotheses. The data that is used to build hypothesis-models can either be raw or derived data or data that is generated from the behaviors of other models or simulations. A key distinctive element of the present approach is to drive hypothesis testing and generation from hypothesis-models that are built from data rather than driving hypothesis testing and generation directly from the data itself. Many methods typically drive hypothesis testing and generation directly from the data. Driving hypothesis testing and generation directly from the data can result in potentially noisier hypotheses due to the increased noise in raw data versus the lower amount of noise in models that are built from the data.

An additional advantage of the method of the present approach lies in the fact that models built from data are typically much smaller in size than the data that they represent. This makes hypothesis testing and generation from models more computationally efficient, especially in large data environments. As the data volume continues to increase rapidly, the scalability of the method of the present approach therefore becomes increasingly valuable.

More generally, data driven hypothesis testing and generation is important in domains where there may not be a priori mathematical models of the underlying system that is being modeled. In many complex, adaptive systems, the relationship between system behavior and the underlying features representing the system can be highly non-linear and multi-dimensional. Modeling these systems with a priori mathematical models from which hypotheses can be tested and generated can lead to significant biases and resulting errors. For these types of applications, empirical hypothesis generation and testing is important, and forms the motivation for the present approach.

To test a hypothesis, the user provides data inputs to the hypothesis-models and Flexscape will produce probability distributions for model outputs. To generate a hypothesis, the user defines desired model output states, and Flexscape will produce states for data inputs that will maximize the probability of achieving the desired output states. The data that is used by Flexscape to test and generate hypotheses can come either from existing databases that contain raw or derived data, or “behavioral” databases that contain data that describe the behaviors of “primary” models or simulations run under different conditions. The hypotheses in the former case represent hypotheses that are based on hypothesis-models built directly from the data; the hypotheses in the latter case represent hypotheses that are based on hypothesis-models that are built from the behaviors of primary models or simulations under different conditions. The primary models or simulations can themselves be derived either from data or from a priori knowledge. Hypotheses based on primary models or simulations that are built from data can be more informative in cases where the underlying data has significant amounts of noise, as these models or simulations may be viewed as noise filters that increase the signal to noise of the data environment.

In addition, filters can be applied to the data coming from raw or derived databases or from behavioral databases prior to hypothesis generation in order to improve the signal to noise of the data environment. The filtered data can be used as the basis for both hypothesis testing and generation resulting in potentially more informative hypotheses. The hypotheses that are generated by Flexscape can also be used in a feedback scheme to refine and focus the data gathering process. If a hypothesis is identified that indicates a particular control strategy is informative, more data can be gathered to further test and validate that strategy. This process can be repeated iteratively to progressively refine and adapt the hypotheses.

The Flexscape system has three core components:

a. Automatic hypothesis-model building from data.
b. Hypothesis testing using the hypothesis-models.
c. Hypothesis generation using the hypothesis-models.

The automatic hypothesis-model building component can work with both complete and incomplete data sets where the incomplete data sets can have missing data fields. One or more models can be built directly from the data. Hypothesis testing generates output predictions from the hypothesis-models given a set of input conditions defined by input features being in specified states. For hypothesis generation, Flexscape uses optimization techniques to generate one or more hypotheses automatically from the hypothesis-models.

In a preferred embodiment of the present approach, the three core components are further implemented as described below:

a. Automatic Hypothesis-Model building from data:

In a preferred embodiment of the present approach, the user can specify the variables in the data that represent "target" variables against which hypotheses are subsequently tested and generated. The user can also specify, either through automated methods or by using human judgment, variables that can be ignored from future consideration. The ability to ignore variables from future consideration becomes important when the number of variables is large. The remaining variables represent "control" variables whose states translate into the hypotheses against the target(s). In the method of the present approach, information theory based measures form the basis for automated feature selection.

In order to improve the computational efficiency of hypothesis-model building, it is often useful to decompose data sets into smaller data subsets. Data sets can be decomposed into one or more data subsets where each data subset contains either a subset of data records ("row subsets") or a subset of features ("feature subsets") or a subset of both data records and features ("row-feature subsets"). In a preferred embodiment of the present approach, data subsets can first be decomposed into row subsets. Measures based on mutual information can then be used to identify informative feature subsets within each row subset to generate a population of smaller row-feature subsets. In a preferred embodiment of the present approach, optimization techniques can be used to guide the selection of the informative feature subsets consistent with user provided constraints. For example, the user might require that an individual feature appear in a predetermined number of feature subsets. The resulting row-feature subsets are used for subsequent hypothesis-model building. One or more hypothesis-models can be automatically generated from each row-feature subset. In the method of the present approach, one or more hypotheses can be generated from individual hypothesis-models, thus providing a plurality of hypotheses that can subsequently be validated. This latter characteristic of the present approach is important in complex systems where some hypotheses may be infeasible to implement.

In a preferred embodiment of the present approach, transparent models such as Bayesian network models or decision tree models are used as the modeling paradigm for building hypothesis-models. Such modeling paradigms provide an explanatory capability that is hard to achieve with black box modeling paradigms such as neural networks. In addition, the use of Bayesian network models facilitates the estimation of missing data values during the hypothesis-model generation process. Furthermore, confidence measures of hypotheses generated from Bayesian models are most directly related to inherent epistemic uncertainty in the data. In other modeling paradigms such as neural networks, the inherent epistemic uncertainty is often confounded with model structure uncertainty resulting in potentially higher bias in the resulting hypotheses.

b. Hypothesis testing using the models:

The population of one or more hypothesis-models generated from the data can be used to test hypotheses against the target variables. Data evidence is presented to a subset of the control variables and the states of the target variables are predicted by the hypothesis models. In a preferred embodiment of the present approach, if data evidence is not presented to a specific control variable, the prior probability distribution for the states of the control variable is used to assign a state for the control variable. In a preferred embodiment of the present approach, this process is repeated multiple times to generate a distribution of target variable predictions. The distribution of target variable predictions can then be analyzed to generate consensus predictions for the target variable(s).

c. Hypothesis generation using the hypothesis-models:

The population of one or more hypothesis-models generated from the data can further be used to generate hypotheses against the target variables. Searching techniques can be used to identify combinations of specific control variable states that maximize the probability of target variables being in desired states. In a preferred embodiment of the present approach, optimization techniques are used to search the control variable state space efficiently in order to generate hypotheses. Further, in a preferred embodiment of the present approach, the LeapWorks Adaptive Optimization Engine is used to search the control variable state space using multiple, diverse optimization methods to generate multiple hypotheses. (J.B. Elad et al., US Patent 5,195,172 issued March 16, 1993, J.B. Elad et al., US Patent 5,428,712 issued June 27, 1995) The application of one or more optimization techniques to search the control variable state space permits the identification of a plurality of hypotheses that satisfy the user defined constraints. In the method of the present approach, statistical confidence measures associated with each hypothesis are automatically generated as outputs.

Overall Process Flow:

In Figure 1, block 104 shows raw or derived data being fed into block 102 where data filtering can be performed using information measures to identify the most informative features. The enriched data set is then fed into block 101 where the hypothesis-models are built. The hypothesis models are then fed into block 100 where hypotheses are generated using optimization techniques and also tested.

In an alternative embodiment of the present approach, either data from block 106 or a priori knowledge from block 108 is fed into block 107 to drive a modeling and simulation engine. Data generated from the simulations is used to populate a behavioral database in block 105. The data from the behavioral database is fed into block 103 where data filtering can be performed using information measures to identify the most informative features. The enriched data set is then fed into block 101 where the hypothesis-models are built. The hypothesis models are then fed into block 100 where hypotheses are generated using optimization techniques and also tested.

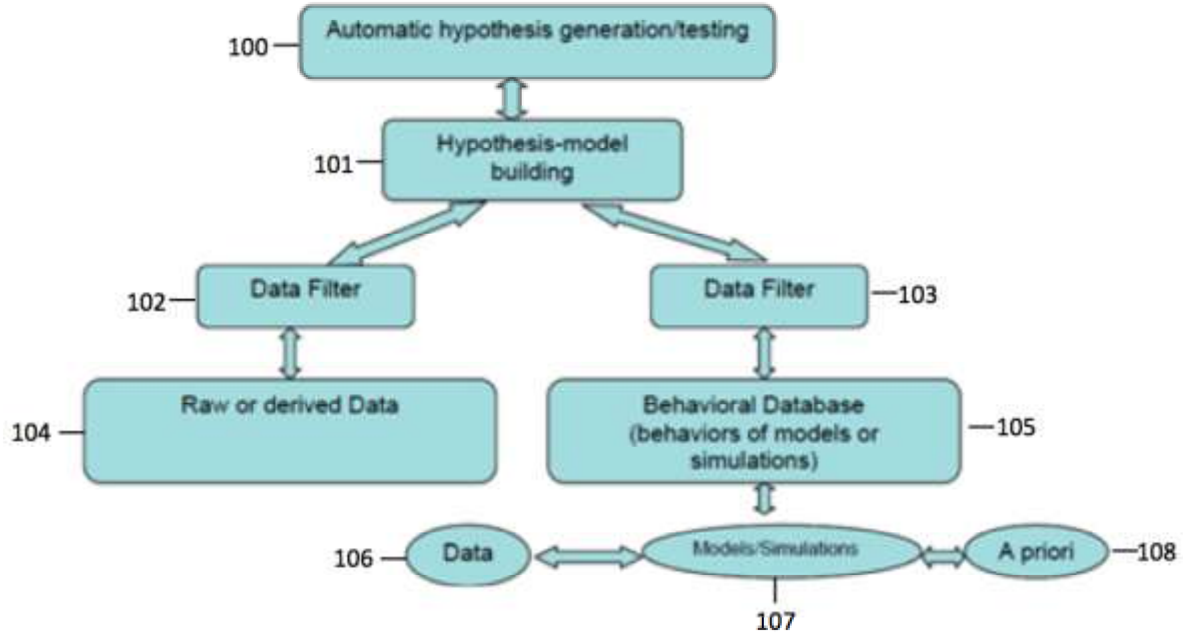
100
Automatic hypothesis generation/testing
101
Hypothesis-model building
102
Data Filter
Data Filter
103
104
Raw or derived Data
Behavioral Database (behaviors of models or simulations)
105
106
Data
Models/Simulations
A priori
108
107

Figure 1. Block diagram of Flexscape system

Examples of applications:

a. Modeling future behaviors from models and simulations of complex, adaptive systems
    - Generate a behavioral data base that encodes future behaviors of models and simulations of complex, adaptive systems such as battlefield environments, infectious and chronic disease spread, manufacturing processes, in the presence of changing input conditions.

- Automatically build a population of behavioral hypothesis-models from the behavioral data that anticipate future behaviors
- Generate and test prognostic hypotheses against the anticipated future behaviors using the behavioral hypothesis-models

b. Generating and testing hypotheses directly from data bases

- Build hypothesis-models directly from existing data bases such as those in health care and life sciences, manufacturing or financial domains.
- Generate and test hypotheses using the hypothesis-models against a range of target variables consistent with potentially changing constraints.

c. Prognostic Hypothesis Generation in dynamic data environments

- The capabilities summarized in bullets (a) and (b) directly above are particularly valuable in the battlefield. From (a), if a sensor system (or sub system) can be modeled as a complex, adaptive system, future behaviors of the system can be simulated under different treatment options. The method of the present approach can analyze a behavioral database that encodes the behavior of such systems under different treatment options to determine the most likely decision options as early as possible. This type of analysis can potentially improve outcomes through early and targeted actions.

Extensions to Dynamic Risk Management:

In a complex, dynamic, data driven environment where uncertainty is the norm, it is essential that principled data analysis techniques be used to both assess and control risk. In this application, we define risk in terms of the probabilistic uncertainty in achieving a desired objective. In particular, we focus on the problem of dynamic risk management where there is a temporal component that must be taken into account. There are many classical approaches to temporal forecasting, including the use of Hidden Markov models, recurrent neural networks, and linear approaches such as ARIMA ("Dynamic Bayesian Networks: Representation, Inference and Learning", Ph.D dissertation**,** Kevin Patrick Murphy, University of California, Berkeley, 2002). These methods often require the user to know in advance the time horizons that can influence a future outcome. Moreover, they cannot always effectively model long term dependencies and do not generally permit the introduction of human domain knowledge. Further, many classical approaches do not deal efficiently or effectively with multivariate inputs and/or outputs.

An effective approach to dynamic risk management that alleviates the problems outlined above is to use a hybrid strategy where human domain expertise can be used to guide an empirical data driven approach to discover the optimal (variable,time) pairs that can influence a future outcome. The method of the present approach describes a multi- stage approach towards implementing such a hybrid strategy:

a. Information theory based discovery of informative time lags in a dynamical data environment:

Each input variable x_i is expanded into a *variable pair* (x_i,t_j) for multiple preceding times t_j that cover an envelope lag period that can be estimated from domain knowledge. The resulting data table can potentially be high dimensional as each input variable is now replicated at multiple time points. The methods of the present approach describe the reduction of the dimensionality of large temporal data sets using information theory. A high dimensional temporal space can be searched efficiently using genetic algorithms or other optimization technologies that use mutual information metrics as the fitness functions to identify key variable pairs that influence the desired target pair ($y,t_{\text{future horizon}}$) at a future time horizon.

The proposed approach can be used in a multi-scale fashion at successive levels of temporal resolution to identify optimal time windows. For example, an initial data table can be created with the temporal unit being weeks; once a set of specific informative week-based lags have been identified, a second data table can be created by resolving the selected week(s) at higher temporal resolution.

An important advantage of the methods of the present approach to temporal pattern discovery lies in the ability to identify *combinations* of temporal patterns that, working together, can influence a target variable at a future time. In complex environments, it is often the case that multiple variables in specific states at different times are informative to influencing a future outcome. The methods of the present approach include the extension of mutual information calculations to multi-dimensional variable sets in a scalable fashion. The critical variable pairs are thus identified in the context of inter-variable interactions in a dynamic environment. A smaller subset of variable pairs that participate most frequently in informative inter-variable interactions can be used to reduce the dimensionality of the data environment in order to build more compact, informative Bayesian network (BN) models as described below.

b. Sorting the selected most informative variable pairs in descending order according to the time lags (from maximum time lags to minimum time lags) to drive a Bayesian network structure learning algorithm such as the well known K2 algorithm:

There are many well known Bayesian network structure learning algorithms described in the literature (see for example “Learning Bayesian Networks”, Richard E. Neapolitan, Prentice Hall Series in Artificial Intelligence, 2003 and references contained therein). Many of the well established methods such as the K2 algorithm assume a given node ordering of the variables that can drive the structure development from root nodes to leaf nodes. The methods of the current approach describe sorting the informative variable pairs identified in step 1 in descending order of time lags to ensure that the leaf nodes within the BN follow earlier nodes from a time sequencing standpoint to preserve causality. This is a key inventive step in the automatic generation of dynamic Bayesian networks.

One or more BN’s can be automatically generated from the data depending on the number of variable pair feature sets that are selected from step 1. The ensemble of Bayesian networks can be scored for quality and a subset of Bayesian networks can be selected as models that can be used to provide risk estimates using probabilistic optimization methods that are outlined below.

c. Applying probabilistic optimization/inductive reasoning on each of the BN's described in step b to generate a sequence of actions that can be taken at preceding times across different control variables to optimally influence the target pair at a future time horizon. This optimization can be performed with multiple temporal/process constraints. Applying optimization techniques on dynamic Bayesian networks represents an important inventive step in this application as a means for enabling dynamic risk control.

d. The dynamic Bayesian networks generated in step b can also be used to forecast risk by performing a forward inference to estimate the likelihood of the (target,time) pair at a future time.

The key inventive step in this application includes the combination of three technology components for enabling scalable dynamic risk assessment and control:

1. Identification of informative (variable, time) pairs against a future (target,time) outcome using an information theory based approach.
2. Automatic generation of dynamic Bayesian networks from the informative pairs described in step 1.
3. Application of optimization methods on the dynamic Bayesian networks to optimally control risk.

Domain Examples for Methods of Present Approach:
Prognostic Situational Awareness:

With the prevalence of new types of sensors penetrating the battlespace environment and the resulting growth in data availability, there are excellent opportunities for modeling the sensor data in a prognostic sense where the target emerges at a future time. The target states can be defined at a level of resolution appropriate to the sensor and the scenario being modeled. Generating and testing sensor driven hypotheses around future threat states represents a critical synergy between the technologies developed by QLI and the fundamental requirements for OA.

In the following section, we present an application of the Flexscape technology in the area of combinatorial chemistry for drug design. This problem has several characteristics that are similar to those that may be encountered in a situational awareness type of scenario:

1. The target state occurs very rarely – e.g., it is a "needle in a haystack".
2. The number of variables is large – this is similar to the case with image sensor data where in principle, every pixel can represent a variable.
3. The data is noisy with many potential confounders that may be typical of complex sensor environments.

Example:
Combinatorial chemistry application/Rational drug discovery:

As an example of the method of the present approach, we present an application from combinatorial chemistry where the objective is to identify combinations of chemical sub-structures that maximize the likelihood that a molecule has the desired biochemical activity against a specified target. Generating hypotheses around optimum sub structures can facilitate new approaches to rational drug discovery. In this example, we use a data set consisting of 7812 compounds where each compound is described by 960 binary structural descriptors. Only 56 compounds are active against the target, with the remaining 7756 compounds inactive. In the method of the present approach, mutual information measures were used to reduce the 960 binary structural descriptors into an initial list of the 100 most informative individual descriptors. Mutual information measures were then used to further reduce the 100 most informative features down to 12 features that participated most often in informative combinations against the target. A Bayesian network was built automatically from the reduced data set (Figure 2). Optimization techniques were then applied to the Bayesian network to maximize the likelihood that the Activity feature is in the active state. The results are summarized in Table 1 below. The four decision features are the parents of the Activity feature, representing the Markov blanket, as shown in Figure 2. The remaining descriptors are denoted as “observable” features. The hypothesis generated by the method of the present approach specifies that all the decision structural features should be present to maximize the probability that the compound is active. Further, probabilities for the remaining features to be present are provided. The overall probability that this hypothesis results in a biochemically active compound is 0.5039, which is significantly enhanced over the 0.0072 baseline probability derived from the data statistics.

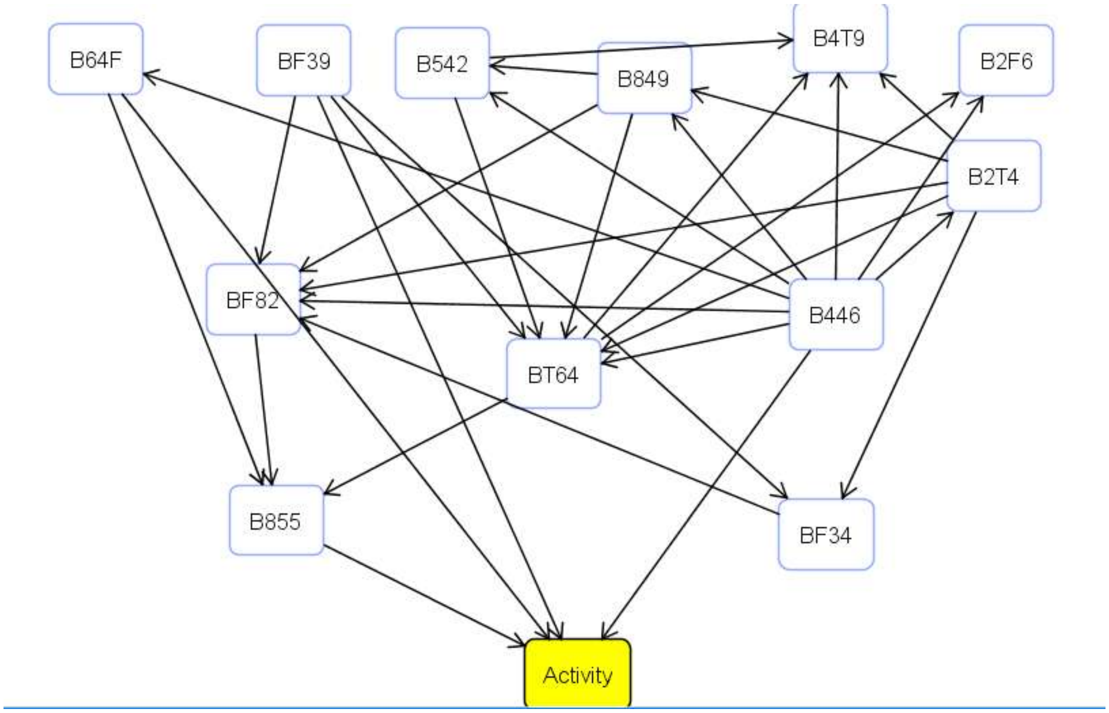

Figure 2. Bayesian Network learned from combinatorial chemistry data set

Descriptor Type	Descriptor	Prob(Absent)	Prob(Present)
Decision	B446	0	1
	B64F	0	1
	B855	0	1
	BF39	0	1
Observable	B2F6	0.1102	0.8898
	B2T4	0.025	0.975
	B4T9	0.0463	0.9537
	B542	0.0417	0.9583
	B849	0.0105	0.9895
	BF34	0.4921	0.5079
	BF82	0.5967	0.4033
	BT64	0.1232	0.8768

Table 1. Hypothesis generated from Bayesian network

4.2 Analysis of Code 30 Data

In this section, we describe the use of QLI's Predictive Analytics component to analyze Code 30 data. We examined a variety of datasets, both classified and unclassified, to try to find interesting predictive analytics problems on which we could use our software. The datasets we examined included sensor data and images, documents including values and features from sensors, notes, records, and information provided before an exercise or kept as part of an exercise. We examined data from two ONR exercises – Green Devil was executed in the summer of 2010 and TNT in November 2010. The data itself was spread across the spectrum from highly unstructured data (chat logs, images, etc.) to very structured data (XML files, spreadsheets, etc.). Although we were able to get some results with data concerning predictions about certain types of events occurring, the majority of the data was not well-suited for our software due to both the type of data and the nature of the scenarios from which the data was derived.

There are two main data characteristics required by QLI's Data Analytics platforms in order to meaningfully process it. First, the data must be very structured. This immediately eliminated a large portion of the data from consideration. Although structure is necessary, it is not sufficient. Firstly, our software does not currently support XML files. That was an inconvenience, but we could have worked around it if not for the second main issue. We seek data that contains a predictive analytics problem. To us, a predictive analytics problem means that the data contains a number of features with one feature in particular representing the target feature. Note that by target feature we mean a feature that we are interested in making predictions on (rather than say the type of target with a bull's-eye). The data should contain many records of the various features including the state of the target feature. These can be used for training purpose and then we can find patterns, generate hypotheses, or make predictions.

In general, even the structured data lacked the type of target feature that we need. Oftentimes the target feature is a sort of "ground-truth," but the nature of the exercise apparently did not lend itself well to this type of analysis. Typically target features may be simple binary features seeking to classify the record as “Good” or “Bad,” “Present” or “Absent,” etc. We are not limited to the binary case; we can handle target states with more than two categorical features and we can also handle continuous target features.

One of the issues with the inclusion of target features is that they typically cannot be obtained directly from sensors. For training data, oftentimes someone with domain knowledge needs to record what the target state is (either manually using domain expertise or through other methods). If someone with domain knowledge of these exercises had added some target features, we might have been able to produce more fruitful results.

That said, we did identify some datasets that seemed to include a target. The datasets were CSV files where the columns represented a number of geographical features and the rows represented different locations. There was also another feature that appeared to represent the likelihood of a particular type of adverse event occurring at that location. We used that likelihood as our target feature and we were able to build models to predict that likelihood with a moderate amount of success. However, it appears as though that likelihood was merely the output of another performer’s analysis that likely included some information (likely domain knowledge) not available to our software. In essence then our software was predicting what that performer would predict, so we don’t believe that to be too interesting of a problem.